

Tema 4: Alineamiento múltiple de secuencias

Introducción a la filoinformática – pan-genómica y filogenómica microbiana, TIB2022, 22-26 enero, 2024 CCG-UNAM, Cuernavaca, Mor. México
<https://github.com/vinuesa/TIB-filoinfo>

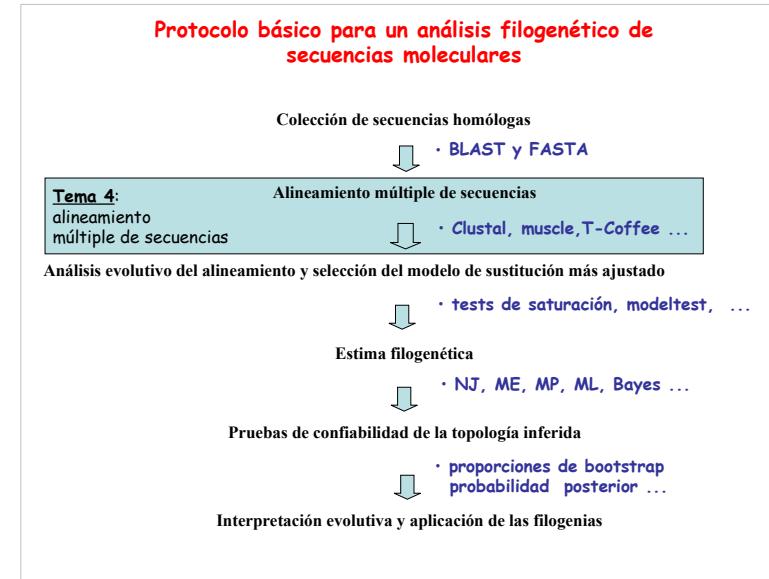
Introducción a la Filoinformática: Pan-genómica y Filogenómica Microbiana – NNB & CCG-UNAM, 22-26 enero 2024

Pablo Vinuesa (vinuesa [at] ccg.unam.mx)
Programa de Ingeniería Genómica, CCG-UNAM, México
<http://www.ccg.unam.mx/~vinuesa/>
[@pvinmex](https://twitter.com/pvinmex)

Todo el material del curso (presentaciones, tutoriales y datos de secuencias) lo encontrarás en:
<https://github.com/vinuesa/TIB-filoinfo>

• Tema 4: Alineamientos múltiples

1. Alineamientos múltiples y el problema de las repeticiones, sustituciones e indeles
2. Alineamientos múltiples progresivos usando programas de la familia Clustal
3. Formatos de secuencia y su interconversión
4. Alineamiento de secuencias codificadoras de proteínas usando RevTrans
5. Alineamiento de genes ribosómicos usando RDP-II y GreenGenes
6. Aln. múltiples usando clustal-omega (clustalo)
7. Aln. múltiples usando mafft
8. Edición y despliegue de alineamientos múltiples con SeaView



- Cualquier estudio de filogenético o de evolución molecular basado en secuencias necesita de un alineamiento múltiple para determinar las correspondencias de homología a nivel de los residuos individuales o caracteres.
- El alineamiento representa una hipótesis sobre la homología de los caracteres
- La mejor manera de representar estas homologías entre caracteres es escribiendo los residuos homólogos en columnas, generándose una matriz de m x n (secs. x posic) residuos, en la que cada columna contiene a residuos o caracteres homólogos

Generación de alineamientos múltiples – consideraciones generales

- El problema de la proteínas multidominio y homología local

Muchas proteínas contienen múltiples dominios, resultado de fusiones de dominios o barajado de exones, entre otros mecanismos. Recuerden que **BLAST hace alineamientos locales** y ordena los HSPs por valores de expectancia decrecientes. De ahí que proteínas que comparten un mismo dominio funcional grande, pero no otros menores, pueden ser recuperados como hits de una misma proteína problema con alto score y bajo valor de expectancia.

Estas proteínas comparten sólo un dominio, es decir presentan homología local pero no global.

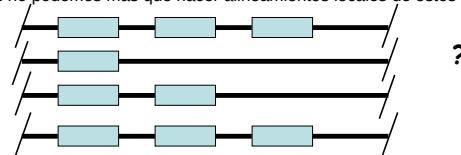
Dado que los algoritmos de alineamiento múltiple generan alineamientos globales, necesitamos recortar los hits de proteínas con diferentes dominios para dejar sólo las regiones que comparten homología local

Generación de alineamientos múltiples – consideraciones generales

• El problema de las repeticiones

Muchas **proteínas multidominio** pueden presentar diverso grado de **repetición de dominios** particulares. Puede llegar a ser muy complejo o imposible hacer el alineamiento global de las proteínas si difieren en el número y orientación de estas regiones repetidas.

A veces no podemos más que hacer alineamientos locales de estos dominios.



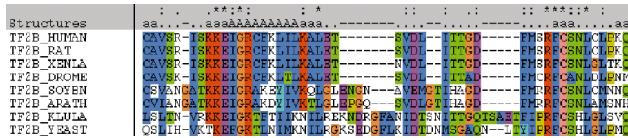
A nivel de DNA se dan también regiones repetidas, muchas veces involucrando a unos poco ntos. como es el caso de los **microsatélites y otras regiones repetidas**. Con frecuencia estas regiones son imposibles de alinear objetivamente. Suelen acumularse en regiones no codificantes del genoma, incluyendo intrones, o en regiones codificantes hipervariables como espaciadores intergénicos transcritos o regiones reguladoras (UTRs).

Este tipo de "repeats" cortos son poco frecuentes a nivel de aminoácidos, si bien a este nivel es común encontrar regiones o **dominios "de gran escala" repetidos**. Un ejemplo clásico de este fenómeno son las **calmodulinas**.

Generación de alineamientos múltiples – consideraciones generales

• El problema de las sustituciones

Es importante recordar que por debajo del **20% de identidad** a nivel de sec. de AA es ya imposible que se pueda obtener un alineamiento múltiple (o pareado) confiable si nos basamos para obtenerlo sólo en la secuencia primaria, ya que entramos en la **zona de penumbra** (saturación mutacional)



Un par de secuencias de ntos al azar presentarán en promedio un 25 % de identidad.

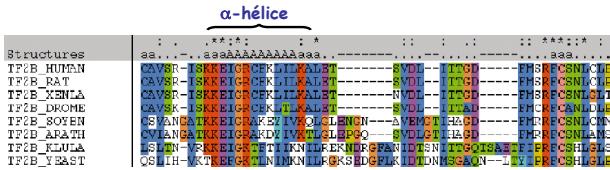
Por tanto, siempre que sea posible, hay que realizar los alineamientos múltiples en base a las secuencias traducidas, es decir, sobre AAs (igual que al hacer búsquedas en bases de datos de secuencia), que son mucho más informativos (20 caracteres vs. 4) y con tasa evolutiva mucho menor.

Generación de alineamientos múltiples – consideraciones generales

• El problema de las sustituciones

Al examinar alns. múltiples de proteínas se observan dos patrones de sustitución:

- Existen bloques de 5 a 20 residuos con alto nivel de identidad y similitud dispersos entre regiones de menor similitud. Estos bloques corresponden típicamente a elementos estructurales como **α-hélices y pliegues β** que evolucionan más lentamente que los loops o bucles que los interconectan

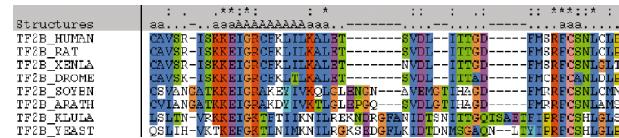


- Las **columnas alineadas** con múltiples estados de carácter **tienden a presentar residuos de características bioquímicas similares** (I, A, V, L; S, T; R, K; etc.). Esta conservación de residuos similares es particularmente patente en los bloques correspondientes a elementos de estructura secundaria, sitios activos o de unión a ligandos.
 La propiedad bioquímica más conservada es la de polaridad/hidrofobicidad.

Generación de alineamientos múltiples – consideraciones generales

• El problema de los indeles (inserciones/delecciones)

Cuando por eventos de inserción o delección (**indels**) las secuencias homólogas presentan distintas longitudes, es necesario introducir "**gaps**" en el alineamiento para mantener la correspondencia entre sitios homólogos situados antes y después de las regiones afectadas por indeles. Estas regiones se identifican mediante guiones (-).



Los "indeles" no se distribuyen aleatoriamente en las secuencias codificadoras (CDSSs).

Casi siempre aparecen **ubicados entre dominios funcionales o estructurales**, preferentemente en bucles (loops) que conectan a dichos dominios.

Esto vale tanto para RNAs estructurales (tRNAs y rRNAs) como para proteínas.

No suelen interrumpir el marco de lectura.

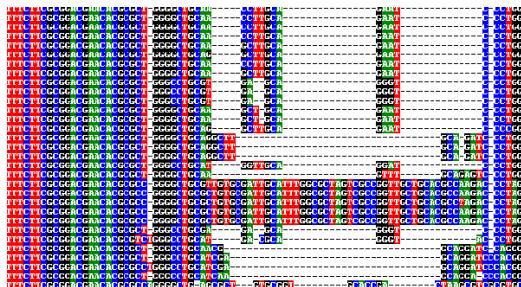
- Generalmente se usan **sistemas de penalización de gaps afines: GP = gap + e(L-1)** en el cálculo del puntaje (score) de un alineamiento múltiple
 [gap = costo apertura de gap; e = costo extensión del indel; L longitud del indel]

Licencia Creative Commons 4.0, no comercial

Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
<https://creativecommons.org/licenses/by-nc/4.0/>

Generación de alineamientos múltiples – consideraciones generales

- A mayor **distancia genética** (evolutiva) entre un par de secuencias, mayor será el número de mutaciones acumuladas. Dependiendo del tiempo de separación de los linajes y la tasa evolutiva del locus, puede llegar a ser imposible alinear ciertas regiones debido a fenómenos de **saturación mutacional y/o acúmulo de indeles**. En loci de evolución muy rápida como intrones o espaciadores intergénicos, los fenómenos de saturación mutacional pueden darse incluso cuando se comparan secuencias de organismos evolutivamente próximos (mismo género o familia), como muestra el ejemplo abajo (ITSs de *Bradyrhizobium* spp).



¡Las regiones de homología dudosa deben de ser excluidas de un análisis filogenético!
 Debemos de maximizar a toda costa la relación entre señal/ruido

Alineamientos múltiples - algoritmos

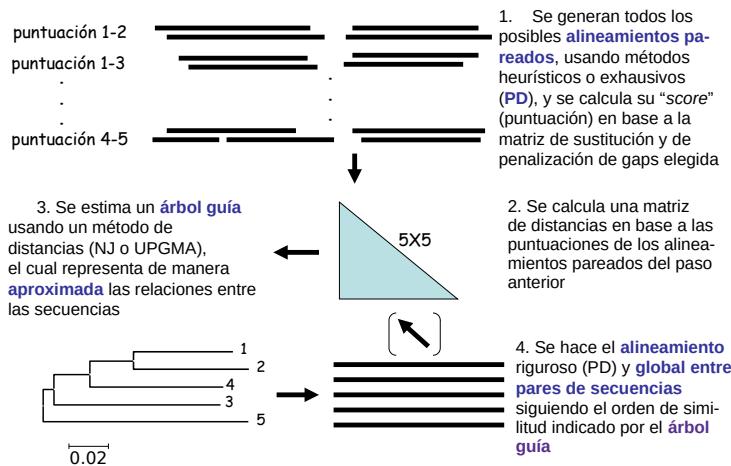
Existen diversas estrategias computacionales para obtener alineamientos múltiples de manera (semi)automática para conjuntos grandes (cientos - miles) de secuencias.

1.- Implementación de algoritmos de alineamiento progresivo.

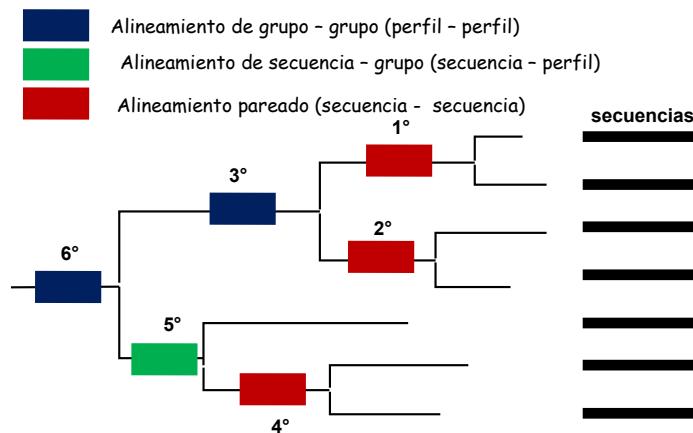
Así como los alns. múltiples son indispensables para reconstruir filogenias a partir de secs, un árbol de relaciones filogenéticas representa información muy valiosa para guiar la generación de un alineamiento múltiple.

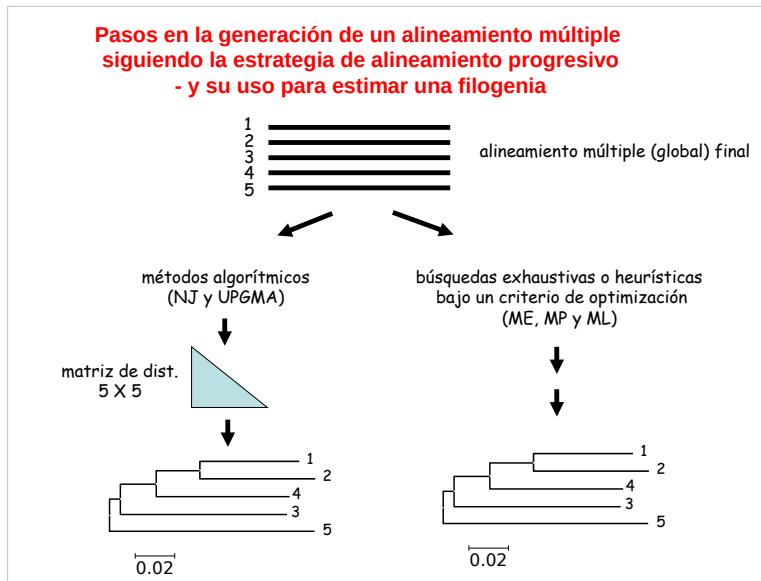
La mayor parte de los alineadores automáticos modernos se basan en este tipo de algoritmos. Construyen un **árbol guía** aproximado a partir de distancias calculadas entre todos los pares posibles de secuencias. De la matriz de distancias resultantes se construye un árbol usando un método algorítmico (**NJ o UPGMA**). El árbol guía resultante se emplea para **construir el alineamiento de manera progresiva**. Las dos secuencias más similares se alinean primero usando PD y una matriz o esquema de ponderación particular. Una vez alineado el primer par, los gaps generados ya no se mueven. Este par es tratado como una sola secuencia y es alineada contra la siguiente secuencia o grupo de secuencias más próximas en el árbol. Se repite el proceso hasta que todas las secs. están alineadas. El proceso es suficientemente rápido como para alinear varios cientos de secuencias. Son menos precisos que los métodos basados en la WSPs, pero muchísimo más rápidos.

Pasos en la generación de un alineamiento múltiple siguiendo la estrategia de alineamiento progresivo



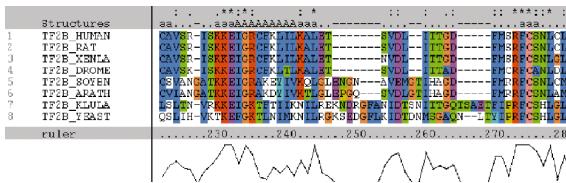
Pasos en la generación de un alineamiento múltiple siguiendo la estrategia de alineamiento progresivo con un árbol guía





Alineamientos múltiples progresivos usando Clustal

- **La familia** Clustal es posiblemente la más popular para hacer AMs de nt y aa
- Existen versiones para todas las plataformas y en red (<http://www.clustal.org/>)
- La primera versión (Clustal) salió en 1988, la última, **Clustal X**, en 2007 (última Vers. = 2.1)
- **ClustalX (X-windows Clustal)** lee secuencias en diversos formatos, calcula un **árbol guía NJ**, usando algoritmos heurísticos o exhaustivos sobre aln. locales basado en **distintas matrices de peso y de penalización de gaps afines y sitio-específicos**. Puede hacer **alineamientos de perfiles** y existen diversas **herramientas de control de calidad del AM**. Permite incluir criterios estructurales para guiar el AM, usando **máscaras estructurales**. Partes del alineamiento o secuencias particulares pueden ser **realineadas** para ir obteniendo un aln global cada vez mejor. Es decir, ClustalX no sólo genera alineamientos (como ClustalW), sino que éstos pueden ser editados y mejorados interactivamente por el usuario. Además, ClustalX (y ClustalW) permite la **reconstrucción y visualización de árboles NJ** y hacer **análisis de bootstrap** sobre los alineamientos. Finalmente, los AMs pueden ser escritos en **diferentes formatos de salida** (CLUSTAL, FASTA, NEXUS, PHYLIP ...)



Alineamientos múltiples progresivos usando Clustal -un ejemplo: alineamiento de GDPs dependientes de NAD

The screenshot shows the ClustalX 1.81 interface. In the top-left, the menu bar includes File, Edit, Alignment, Trees, Colors, Quality, and Help. The "Multiple Alignment Mode" dropdown is set to "Profile Alignment Mode". The "Font Size" is set to 10. The "Multiple Alignment Mode" dropdown is expanded, showing options like "Multiple Alignment Mode", "Profile Alignment Mode", and "Protein Alignment Mode".

In the center, a list of files for alignment is shown, including:

- 1. gpdadrose
- 2. gpdhuanas
- 3. gdakouse
- 4. gpdarabit
- 5. gpdoxcel
- 6. GDP_eucar_ALN
- 7. GDP_eucar_UALN
- 8. Arbol_Haezel
- 9. BayesInference.net.course
- 10. Clase1_Conceptos_basicos_fisiogenet_evol_mol
- 11. Clase2_alineamientos
- 12. Clase2_fn_intro_Aln_pareados
- 13. GDP_AA_UALN

At the bottom, there is a file selection dialog titled "Cuso_Lic_CG_BEG_IV_05" with a list of files and a "Nombre de archivo:" field containing "6_GDP_eucar_UALN".

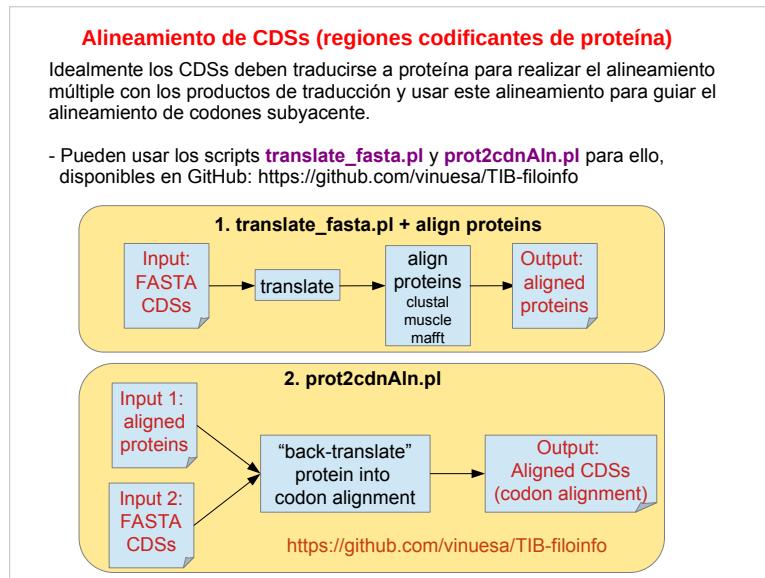
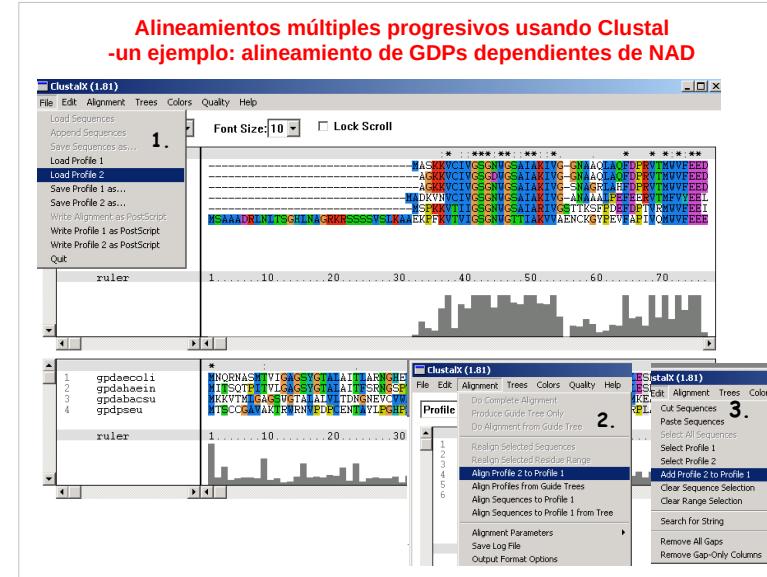
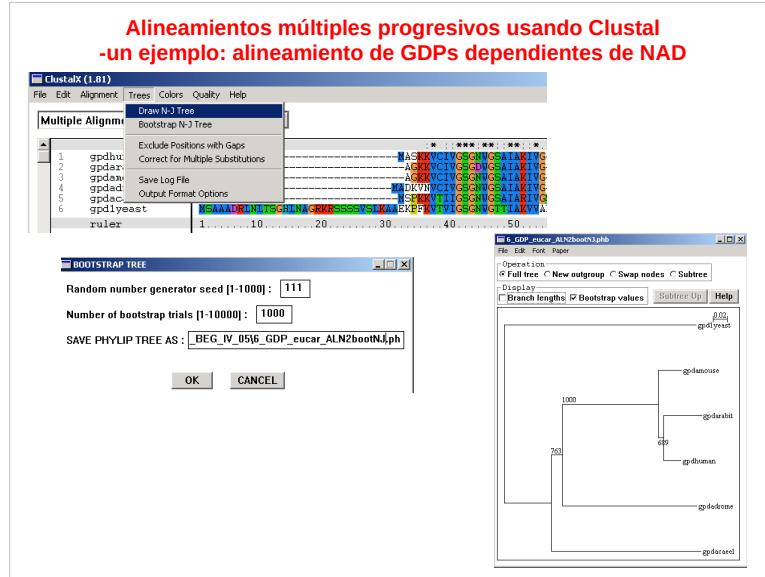
Alineamientos múltiples progresivos usando Clustal -un ejemplo: alineamiento de GDPs dependientes de NAD

The screenshot shows the ClustalX 1.81 interface with the alignment window open. The menu bar includes File, Edit, Alignment, Trees, Colors, Quality, and Help. The "Multiple Alignment Mode" dropdown is expanded, showing options like "Do Complete Alignment", "Produce Guide Tree Only", and "Do Alignment from Guide Tree".

The alignment window displays a multiple sequence alignment of 10 sequences (1-10) with colored residue blocks. To the right, a phylogenetic tree is shown with nodes labeled 1 through 10. Below the alignment, there are two profile plots: one for the top sequence and one for the bottom sequence.

Tema 4: Alineamiento múltiple de secuencias

Introducción a la filoinformática – pan-genómica y filogenómica microbiana, TIB2022, 22-26 enero, 2024 CCG-UNAM, Cuernavaca, Mor. México
<https://github.com/vinuesa/TIB-filoinfo>



Servidores para alinear nts. en base a un alineamiento de proteínas

<http://www.cbs.dtu.dk/services/RevTrans/>

The screenshot shows the RevTrans 1.4 Server interface. At the top, there's a navigation bar with links to CENTERFOR, EVENTS, NEWS, RESEARCH GROUPS, CBS PREDICTION SERVERS, CBS DATA SERVER, PUBLICATIONS, and BIOINFORMATICS EDUCATION PROGRAM. Below the navigation is a main form with sections for 'Instructions', 'Output format', 'Background', 'Software download', and 'Article abstract'. There are fields for 'Paste in DNA sequences' and 'Optional: Paste in peptide alignment', as well as 'Upload file containing DNA sequences' and 'Optional: Upload peptide alignment'. A note states: 'Valid formats: FASTA, MSF and ALN (Clustal) - any gaps will be removed from DNA sequences'. At the bottom are 'Submit query', 'Clear fields', and 'Translate only' buttons.

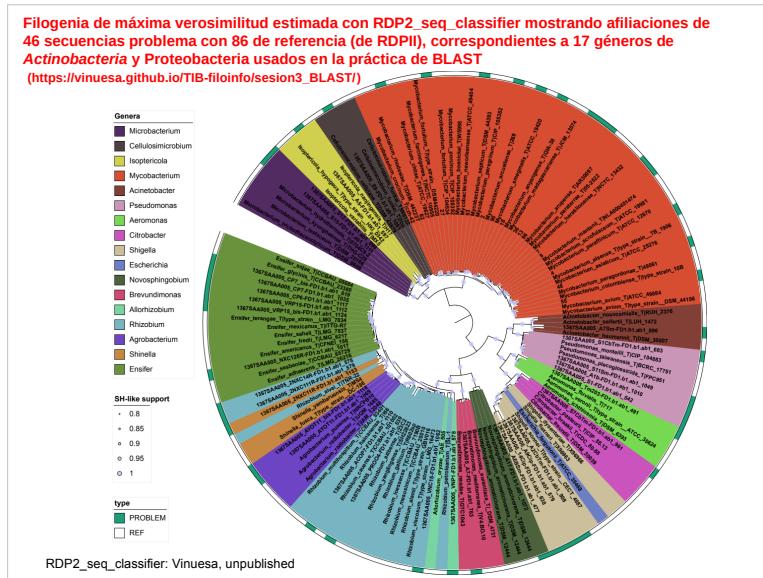
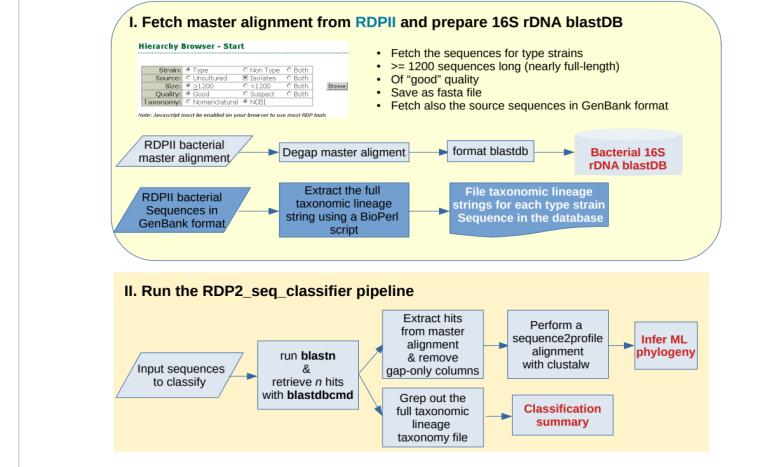
Servidores para alinear secuencias de rRNAs o rDNAs

- Los genes ribosomales representan un problema muy particular en el contexto de alineamientos múltiples. Deben de guiarse usando máscaras de información estructural.

Servidores como **arb-silva** y **RDP-II** proveen herramientas muy útiles en este contexto. Si quieras ver unos tutoriales sobre el uso de estos servidores, visita mi sitio web y busca bajo phylogeny tutorials:
https://www.ccg.unam.mx/~vinuesa/Using_the_GreenGenes_and_RDP-II_servers.html

<https://www.arb-silva.de/>
<http://rdp.cme.msu.edu/>

RDP2_seq_classifier: una tubería bioinformática para alinear, clasificar y analizar filogenéticamente secuencias de rRNAs bacterianos



Formatos de secuencias I) FASTA

- Existen una gran cantidad de estilos o formatos de presentación de secuencias. Muchos programas de análisis filogenético usan su propio formato (Phylip, Nexus, Mega ...)
- El formato más sencillo es el **FASTA**, en el que cada secuencia se identifica mediante un renglón descriptor que comienza con **>** en el siguiente renglón comienza la secuencia

```

>lcl|1 R._galegae
CCGCTGGTCACCTCCGGCAAGCGCGCCATCCACCAGGAAGCGCCTTCCA
CGTGATCAGTCGACCGAAGGCCAGATCTGGTCACGGCATCAAGTCG

>lcl|2 M._plurifarium
CCGGTCGACGCCGTCAGCTGCGTGCATCCACCAGCGGCTCCGCTTA
TGTGACCGTCGACCGAAGGCCAGATCTGGTACCGGCATCAAGGTC

>lcl|3 B._japonicum
CCGGTCAAGTCGGAAGGCCCTGCGCGCCATCCACCAGGAAGCGGCCACTA
CACCGACCAAGTCCACCGAACGCTGAAATTCTCGTACCGGCATCAAGGTC

```

Formatos de secuencias

II) PHYLIP

- **Phylip (interleaved):** no. seqs, no. caracteres
nombre secuencias (máx 10 caracteres) espacio, secuencia ...

```
3      100
R._galegae  CGCGUGGUCA  CCUCGGCAA  GCGCGCAUC  CACCAAGGAAG  CGCCUUCUA
M._plurifa  ...G.C.A.G  ..GU..AGCU  ....U.....  .....CCG  .U..GG...
B._japonic  ...G.CAAGU  .GGAA..CU  .....  .....  .....GA...
CGUCGAUCAG  UCGACCGAAG  GCCAGAUCCU  GGUCACCGGC  AUCAAGGUUC
U.....C....  ....G....  CG.....  ...U.....  .....UC
.AC...C....  ..C.....  CUG.A..U..  C.....  .....
```

- **Phylip (sequential or non-interleaved)**

```
3 100
R._galegae  CCGCTGGTCA  CCTCCGGCAA  GCGGCCATC  CACCAAGGAAG  CGCCTTCCTA
CGTCGATCG  TCGACCGAAG  GCCAGATCTT  GGTCAACCGGC  ATCAAGGTGCG
M._plurifa  CCGGTGAGCG  CCGTCGAGCT  GCGTGCATC  CACCAAGCCGG  CTCCGGCTTA
TGTCGACCAAG  TCGACCGAAG  CGCAGATCTT  GGTACCGGC  ATCAAGGTTC
B._japonic  CCGGTCAAGT  CGGAAGGCT  GCGGCCATC  CACCAAGGAAG  CGCCGACCTA
CACCGACCAG  TCCACCGAAG  CTGAAATTCT  CGTCACCGGC  ATCAAGGTGCG
```

Formatos de secuencias

III) NEXUS

```
#NEXUS
[OJO!!!, no usar guiones- (reservado para gaps!), sólo guiones bajos_]

BEGIN TAXA;      [taxa block]
DIMENSIONS NTAX=3;
TAXLABELS
R._galegae;
M._plurifarium;
B._japonicum;
END;

BEGIN CHARACTERS;   [character block]
DIMENSIONS NCHAR=100;
FORMAT DATATYPE=DNA MISSING=? GAP=- MATCHCHAR=. INTERLEAVE=yes ;
MATRIX
[          10    20    30    40    50
[          *     *     *     *     *
R._galegae  CCGCTGGTCACTCCGGCAAGGGCGCCATCCACCAAGGAAGCGCCTTCCTA
M._plurifarium ...G.C.A.G..GT..AGCT..T.....CCG..T..GG...
B._japonicum ...G.CAAGT.GGAA..CT.....GA...

[          60    70    80    90    100
[          *     *     *     *     *
R._galegae  CGTCGATCAAGTCGACCGAAGGGCAGATCCTGGTCACCGGCATCAAGGTCG
M._plurifarium T.....C.....G..CG.....T.....TC
B._japonicum ..AC...C....CTG.A..T..C.....
;
END;
```

Formatos de secuencias: su interconversión

- Cuando preparamos un fichero con nuestras propias secuencias generalmente lo más adecuado es hacerlo en formato FASTA
- Si necesitamos pasarlo a otro formato, una buena posibilidad es hacerlo con [ReadSeq](#)

<http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi>

ReadSeq reconoce automáticamente el formato de entrada y si se trata de aas o nts

- Otra alternativa es escribir un sencillo script de [Perl](#) que haga uso de los objetos y métodos del módulo Bio::AlignIO de [BioPerl](#) (<http://www.bioperl.org>) para interconvertir Formatos ... veremos un ejemplo más adelante, implementado en el script:

[convert_align_format_batch_bp.pl](https://github.com/vinuesa/TIB-filoinfo) disponible en <https://github.com/vinuesa/TIB-filoinfo>

Alineamientos múltiples progresivos usando Clustalw

Desde el directorio en el que están las secuencias llamar a ClustalW

```
vinuesa@teide:~/cd Cursos/BGEIV-06/
vinuesa@teide:~/Cursos/BGEIV-06> ls
ClustalW_cmmds.txt myoglobins.fasta
vinuesa@teide:~/Cursos/BGEIV-06> clustalw
```

```
*****
***** CLUSTAL W (1.83) Multiple Sequence Alignments *****
*****
```

1. Sequence Input From Disc
 2. Multiple Alignments
 3. Profile / Structure Alignments
 4. Phylogenetic trees
- S. Execute a system command
 - H. HELP
 - X. EXIT (leave program)

Your choice:1

Alineamientos múltiples progresivos usando Clustalw

¿No se ve nada conveniente, verdad?, ¿pero qué tal ésto?:

```
clustalw -infile=myoglobins.fasta -align -pwmatrix=blosum -pwgapopen=12
-pwgapext=0.2 -matrix=blosum -gapopen=12 -gapext=0.2 -outorder=aligned
-convert -outfile=myoglobins_align.phy -output=phylip
```

- Y si tenemos muchos archivos fasta de proteína para alinear (extensión .faa), podemos escribir una simple línea de shell como la siguiente para alinearlos todos consecutivamente usando los parámetros por defecto de clustalw:

```
for file in *.faa; do clustalw -infile=$file -align -convert -output=fasta; done
```

- Las opciones de clustal en línea de comando se ven así:

clustalw --options

y la ayuda así:

clustalw --help # puede no funcionar si el documento de ayuda no está en el path


Clustal Omega
 "The last alignment program you'll ever need"

Home
Webservers
Download
Documentation
Contact
News

Introduction

Clustal Omega is the latest addition to the Clustal family. It offers a significant increase in scalability over previous versions, allowing hundreds of thousands of sequences to be aligned in only a few hours. It will also make use of multiple processors, where present. In addition, the quality of alignments is superior to previous versions, as measured by a range of popular benchmarks.

Please note that Clustal Omega is currently a command line-only tool.

A full description of the algorithms used by Clustal Omega is available in the Molecular Systems Biology paper *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega*. Latest additions to Clustal Omega are described in *Clustal Omega for making accurate alignments of many protein sciences*.

Webservers

Clustal Omega can be run online at the following websites (this list will expand in the coming months).

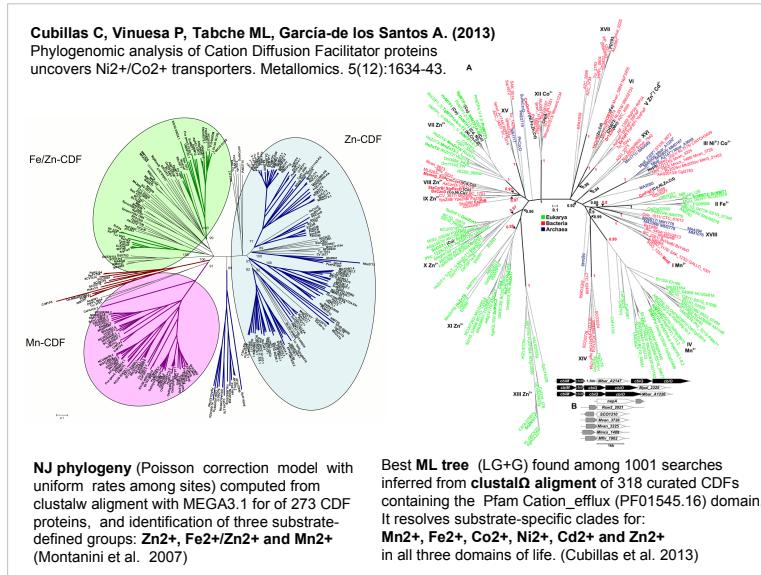
Download Clustal Omega

[Source code](#)

- [Source code .tar.gz \(1.2.4\)](#)

Citing Clustal

1. Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7:539 doi:10.1038/msb.2011.75
2. Sievers F, Higgins DG (2018) Clustal Omega for making accurate alignments of many protein sciences. *Protein Sci* 27:135-145




Clustal Omega
 "The last alignment program you'll ever need"

clustalo --help
Clustal Omega - 1.2.4 (AndreaGiacomo)

Check <http://www.clustal.org> for more information and updates.

- Usage:
 clustalo [-hv] [-i <file>...]
 [-hmm-in=<file>]... [-hmm-batch=<file>]
 [-dealign] [-profile1=<file>] [-profile2=<file>]
 [-is-profile] [-t {Protein, RNA, DNA}] [-infmt={a2m=fasta,clustal,msf,phy[tip],selex,st[ockholm],vie[nrna]}]
 [-distrmat=<file>] [-distrmat-out=<file>] [-guidetree-in=<file>] [-guidetree-out=<file>] [-pileup] [-full] [-full-iter]
 [-cluster-size=<n>] [-clustering-out=<file>] [-trans=<n>] [-posterior-out=<file>] [-use-kimura] [-percent-id]
 [-o <file>] [-outfmt={(a2m=fasta,clustal,msf,phy[tip],selex,st[ockholm],vie[nrna])} [-residue-number] [-wrap=<n>]
 [-output-order={input-order,tree-order}]
 [-iterations=<n>] [-max-guidetree-iterations=<n>] [-max-hmm-iterations=<n>]
 [-maxnumseq=<n>] [-maxseqlen=<n>] [-auto]
 [-threads=<n>] [-pseudo=<file>] [-i <file>]
 [-version] [-long-version] [-force] [-MAC-RAM=<n>]
 - A typical invocation would be:

```
clustalo -i my-in-seqs.fa -o my-out-seqs.fa
```

```
for f in *.fa; do clustalo -i $f -o ${f%.}_cluoAln.fa; done
```

Tema 4: Alineamiento múltiple de secuencias

Introducción a la filoinformática – pan-genómica y filogenómica microbiana, TIB2022, 22-26 enero, 2024 CCG-UNAM, Cuernavaca, Mor. México
<https://github.com/vinuesa/TIB-filoinfo>

MAFFT version 7
 Multiple alignment program for amino acid or nucleotide sequences

Downloaded version
 Build 5.8
 Windows
 Linux
 Source
 Data sources
 Alignments
 mafft-align
 Global
 Local
 Multiple
 Algorithms
 Tree
 Benchmarks
 Feedback

Algorithms and parameters (unfinished)
 MAFFT offers various multiple alignment strategies. They are classified into three types, (a) the progressive method, (B) the iterative refinement method with the WSP score, and (c) the iterative refinement method using both the WSP and consistency scores. In general, there is a tradeoff between speed and accuracy. The order of speed is a > b > c, whereas the order of accuracy is a < b < c. The results of benchmarks can be seen [here](#). The following are the detailed procedures.

(a) FFT-NS-1, FFT-NS-2 – Progressive methods
 Distance matrix based on the number of shared 6-tuples
 Constructing guide tree
 Progressive alignment
 Reconstructing guide tree
 Alignment
 FFT-NS-1
 FFT-NS-2

These are simple progressive methods, i.e., [ClustalW](#). By using the several new techniques described below, these options can align a large number of sequences (up to ~5,000) on a standard desktop computer. The qualities of the resulting alignments are very good. The detailed descriptions are described in Kihara et al. (2002).

MAFFT v7.505 (2022/Apr/10)
<https://mafft.cbrc.jp/alignment/software/>
 MBE 30:772-780 (2013), NAR 30:3059-3066 (2002)

High speed:
 % maftt in > out
 % maftt --retree 1 in > out (fast)

High accuracy (for <~200 sequences x <~2,000 aa/nt):
 % maftt --maxiterate 1000 --localpair in > out (% linsi in > out is also ok)
 % maftt --maxiterate 1000 --genepair in > out (% einsi in > out)
 % maftt --maxiterate 1000 --globalpair in > out (% ginsi in > out)

If unsure which option to use:
 % maftt --auto in > out

--op #: Gap opening penalty, default: 1.53
 --ep #: Offset (works like gap extension penalty), default: 0.0
 --maxiterate #: Maximum number of iterative refinement, default: 0
 --clustalout: Output: clustal format, default: fasta
 --reorder: Outorder: aligned, default: input order
 --quiet: Do not report progress
 --thread #: Number of threads (if unsure, --thread -1)
 --dash: Add structural information (Rozewicki et al, submitted)

SeaView: paquete multiplataforma para alineamientos y análisis filogenético

<https://doua.prabi.fr/software/seaview>



PRABI-Doua
 Pôle Rhône-Alpes de Bioinformatique

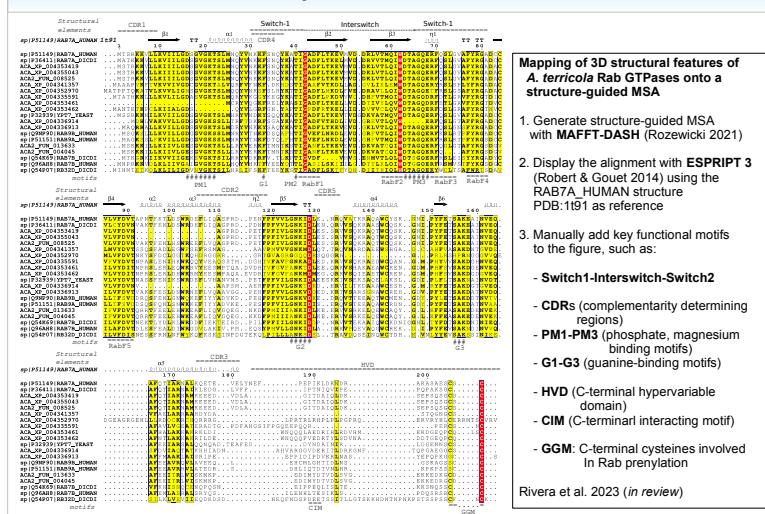
SeaView - Multiplatform GUI for molecular phylogeny

Version 5.0.5

NEW: seaview performs reconciliation between gene and species trees using [TreeRecs](#) version 1.2
 NEW: bootstrap support optionally with the "Transfer Bootstrap Expectation" method
 NEW: trimming-rules to shorten long sequence names in phylogenetic trees
 NEW: 64-bit version for the MS Windows platform
 NEW: multiple-tree windows
 NEW: seaview uses [PhyML 3.6.9](#) to compute parsimony trees
 NEW: seaview runs without GUI using a command line
 NEW: seaview drives the [PhyML v3.1](#) program to compute maximum likelihood phylogenetic trees.
 NEW: seaview drives the [Gblocks](#) program to select blocks of conserved sites.

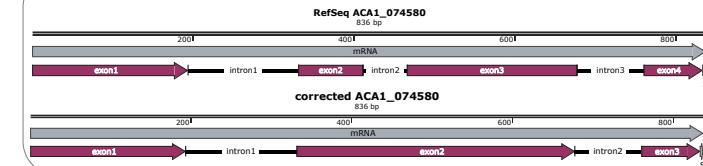
SeaView is a multiplatform, graphical user interface for multiple sequence alignment and molecular phylogeny.

Structure-guided MSA of *A. terricola* Rab7 paralogues and Swiss-Prot reference sequences displayed with ESPRT3 reveals that 75% of them are non-canonical Rabs, lacking different conserved functional motifs and major structural features Rivera et al. 2024. *Microbiology Spectrum (in press)*

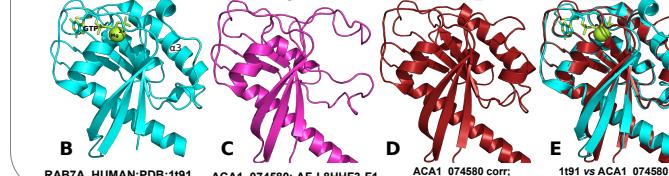


Miniprot alignment of RAB7A_HUMAN protein to the *A. terricola* ACA1_074580 gene and structural modeling of the corrected conceptual product confirms that it encodes for the amoebal Rab7A protein Rivera et al. 2024. *Microbiology Spectrum (in press)*

1. miniprot alignment of RAB7A_HUMAN against ACA1_074580 uncovers a misplaced intron



2. 3D structure for RAB7A_HUMAN (B) and AF models for ACA1_074580 (C), miniprot-corrected ACA1_074580 (D) and alignment of corrected ACA1_074580 on RAB7A_HUMAN (E)



1. Miniprot alignment of RAB7A_HUMAN protein to the ACA1_074580 gene uncovers a misplaced intron in RefSeq annot.
2. Model the corrected ACA1_074580 translation product with AlphaFold2 using a ColabFold notebook (Mirdita 2022)
3. Evaluate TM-scores with TM-align. Display protein alignments with PyMol.